VI Semester

DATA SCIENCE AND ITS APPLICATIONS			
Course Code	21AD62	CIE Marks	50
Teaching Hours/Week (L:T:P: S)	3:0:2:0	SEE Marks	50
Total Hours of Pedagogy	40 T + 20 P	Total Marks	100
Credits	04	Exam Hours	03

Course Learning Objectives:

CLO 1. Demonstrate the proficiency with statistical analysis of data to derive insight from results and interpret the data findings visually

CLO 2. Utilize the

CLO 3. skills in data management by obtaining, cleaning and transforming the data.

CLO 4. Make use of machine learning models to solve the business-related challenges

CLO 5. Experiment with decision trees, neural network layers and data partition.

CLO 6. Demonstrate how social clustering shape individuals and groups in contemporary society.

Teaching-Learning Process (General Instructions)

These are sample Strategies, which teacher can use to accelerate the attainment of the various course outcomes.

1. Lecturer method (L) does not mean only traditional lecture method, but different type of teaching methods may be adopted to develop the outcomes.
2. Show Video/animation films to explain functioning of various concepts.
3. Encourage collaborative (Group Learning) Learning in the class.
4. Ask at least three HOTS (Higher order Thinking) questions in the class, which promotes critical thinking.
5. Adopt Problem Based Learning (PBL), which fosters students' Analytical skills, develop thinking skills such as the ability to evaluate, generalize, and analyze information rather than simply recall it.
6. Topics will be introduced in a multiple representation.
7. Show the different ways to solve the same problem and encourage the students to come up with their own creative ways to solve them.
8. Discuss how every concept can be applied to the real world - and when that's possible, it helps improve the students' understanding.

Module-1: Introduction

What is Data Science? Visualizing Data, matplotlib, Bar Charts, Line Charts, Scatterplots, **Linear Algebra,** Vectors, Matrices, **Statistics,** Describing a Single Set of Data, Correlation, Simpson's Paradox, Some Other Correlational Caveats, Correlation and Causation, **Probability,** Dependence and Independence, Conditional Probability, Bayes's Theorem, Random Variables, Continuous Distributions, The Normal Distribution, The Central Limit Theorem.

Chapters 1, 3, 4, 5 and 6

Laboratory Component:

1. Installation of Python/R language, Visual Studio code editors can be demonstrated along with Kaggle data set usage.
2. Write programs in Python/R and Execute them in either Visual Studio Code or PyCharm Community Edition or any other suitable environment.
3. A study was conducted to understand the effect of number of hours the students spent studying on their performance in the final exams. Write a code to plot line chart with number of hours spent studying on x-axis and score in final exam on y-axis. Use a red '*' as the point character, label the axes and give the plot a title.

Number of hrs spent studying (x)	10	9	2	15	10	16	11	16
Score in the final exam (0 – 100) (y)	95	80	10	50	45	98	38	93

4. For the given dataset mtcars.csv (www.kaggle.com/ruiromanini/mtcars), plot a histogram to check the frequency distribution of the variable 'mpg' (Miles per gallon)

Teaching-Learning Process	1. Demonstration of different charts 2. PPT Presentation for Theorems and different distributions 3. Live coding and execution for visualization with simple examples

Module-2: Hypothesis and Inference

Statistical Hypothesis Testing, Example: Flipping a Coin, p-Values, Confidence Intervals, p-Hacking, Example: Running an A/B Test, Bayesian Inference, **Gradient Descent,** The Idea Behind Gradient Descent Estimating the Gradient, Using the Gradient, Choosing the Right Step Size, Using Gradient Descent to Fit Models, Minibatch and Stochastic Gradient Descent, **Getting Data,** stdin and stdout, Reading Files, Scraping the Web, Using APIs, Example: Using the Twitter APIs, **Working with Data,** Exploring Your Data, Using NamedTuples, Dataclasses, Cleaning and Munging, Manipulating Data, Rescaling, An Aside: tqdm, Dimensionality Reduction.

Chapters 7, 8, 9 and 10

Laboratory Component:

1. Consider the books dataset BL-Flickr-Images-Book.csv from Kaggle (https://www.kaggle.com/adeyoyintemidayo/publication-of-books) which contains information about books. Write a program to demonstrate the following.

- Import the data into a DataFrame
- Find and drop the columns which are irrelevant for the book information.
- Change the Index of the DataFrame
- Tidy up fields in the data such as date of publication with the help of simple regular expression.
- Combine str methods with NumPy to clean columns

Teaching-Learning Process	1. Demonstration of Hypothesis test. 2. PPT Presentation to explore and manipulate data. 3. Live coding of concepts with simple examples 4. Case Study: Extraction of data from Books dataset

Module-3: Machine Learning

Modeling, What Is Machine Learning?, Overfitting and Underfitting, Correctness, The Bias-Variance Tradeoff, Feature Extraction and Selection, **k-Nearest Neighbors,** The Model, Example: The Iris Dataset, The Curse of Dimensionality, **Naive Bayes,** A Really Dumb Spam Filter, A More Sophisticated Spam Filter, Implementation, Testing Our Model, Using Our Model, **Simple Linear Regression,** The Model, Using

Gradient Descent, Maximum Likelihood Estimation, **Multiple Regression,** The Model, Further Assumptions of the Least Squares Model, Fitting the Model, Interpreting the Model, Goodness of Fit, Digression: The Bootstrap, Standard Errors of Regression Coefficients, Regularization, **Logistic Regression,** The Problem, The Logistic Function, Applying the Model, Goodness of Fit, Support Vector Machines.

Chapters 11, 12, 13, 14, 15 and 16

Laboratory Component:

1. Train a regularized logistic regression classifier on the iris dataset (https://archive.ics.uci.edu/ml/machine-learning-databases/iris/ or the inbuilt iris dataset) using sklearn. Train the model with the following hyperparameter C = 1e4 and report the best classification accuracy.
2. Train an SVM classifier on the iris dataset using sklearn. Try different kernels and the associated hyperparameters. Train model with the following set of hyperparameters RBF-kernel, gamma=0.5, one-vs-rest classifier, no-feature-normalization. Also try C=0.01,1,10C=0.01,1,10. For the above set of hyperparameters, find the best classification accuracy along with total number of support vectors on the test data

Teaching-Learning Process	1. Demonstration of Models 2. PPT Presentation for techniques 3. Live coding of all concepts with simple examples

Module-4: Decision Trees

What Is a Decision Tree?, Entropy, The Entropy of a Partition, Creating a Decision Tree, Putting It All Together, Random Forests, **Neural Networks,** Perceptrons, Feed-Forward Neural Networks, Backpropagation, Example: Fizz Buzz, **Deep Learning,** The Tensor, The Layer Abstraction, The Linear Layer, Neural Networks as a Sequence of Layers, Loss and Optimization, Example: XOR Revisited, Other Activation Functions, Example: Fizz Buzz Revisited, Softmaxes and Cross-Entropy, Dropout, Example: MNIST, Saving and Loading Models, **Clustering,** The Idea, The Model, Example: Meetups, Choosing k, Example: Clustering Colors, Bottom-Up Hierarchical Clustering

Chapters 17, 18, 19 and 20

Laboratory Component:

1. Consider the following dataset. Write a program to demonstrate the working of the decision tree based ID3 algorithm.

Price	Maintenance	Capacity	Airbag	Profitable
Low	Low	2	No	Yes
Low	Med	4	Yes	Yes
Low	Low	4	No	Yes
Low	Med	4	No	No
Low	High	4	No	No
Med	Med	4	No	No
Med	Med	4	Yes	Yes
Med	High	2	Yes	No
Med	High	5	No	Yes
High	Med	4	Yes	Yes
high	Med	2	Yes	Yes
High	High	2	Yes	No
high	High	5	yes	Yes

2. Consider the dataset spiral.txt (https://bit.ly/2Lm75Ly). The first two columns in the dataset corresponds to the co-ordinates of each data point. The third column corresponds to the actual cluster label. Compute the rand index for the following methods:

- K – means Clustering
- Single – link Hierarchical Clustering
- Complete link hierarchical clustering.
- Also visualize the dataset and which algorithm will be able to recover the true clusters.

Teaching-Learning Process	1. Demonstration using Python/ R Language 2. PPT Presentation for decision tree, Neural Network, Deep learning and clustering 3. Live coding for the concepts with simple examples 4. Project Work: Algorithm implementation

Module-5: Natural Language Processing

Word Clouds, n-Gram Language Models, Grammars, An Aside: Gibbs Sampling, Topic Modeling, Word Vectors, Recurrent Neural Networks, Example: Using a Character-Level RNN, **Network Analysis,** Betweenness Centrality, Eigenvector Centrality, Directed Graphs and PageRank, **Recommender Systems,** Manual Curation, Recommending What's Popular, User-Based Collaborative Filtering, Item-Based Collaborative Filtering, Matrix Factorization.

Chapters 21, 22 and 23

Laboratory Component:

Mini Project – Simple web scrapping in social media

Teaching-Learning Process	1. Demonstration of models 2. PPT Presentation for network analysis and Recommender systems 3. Live coding with simple examples

Course outcome (Course Skill Set)

At the end of the course the student will be able to:

CO 1. Identify and demonstrate data using visualization tools.
CO 2. Make use of Statistical hypothesis tests to choose the properties of data, curate and manipulate data.
CO 3. Utilize the skills of machine learning algorithms and techniques and develop models.
CO 4. Demonstrate the construction of decision tree and data partition using clustering.
CO 5. Experiment with social network analysis and make use of natural language processing skills to develop data driven applications.

Assessment Details (both CIE and SEE)

The weightage of Continuous Internal Evaluation (CIE) is 50% and for Semester End Exam (SEE) is 50%. The minimum passing mark for the CIE is 40% of the maximum marks (20 marks). A student shall be deemed to have satisfied the academic requirements and earned the credits allotted to each subject/ course if the student secures not less than 35% (18 Marks out of 50) in the semester-end examination (SEE), and a minimum of 40% (40 marks out of 100) in the sum total of the CIE (Continuous Internal Evaluation) and SEE (Semester End Examination) taken together

Continuous Internal Evaluation:

Three Unit Tests each of **20 Marks (duration 01 hour**)

1. First test at the end of 5th week of the semester
2. Second test at the end of the 10th week of the semester
3. Third test at the end of the 15th week of the semester

Two assignments each of **10 Marks**

4. First assignment at the end of 4th week of the semester
5. Second assignment at the end of 9th week of the semester

Practical Sessions need to be assessed by appropriate rubrics and viva-voce method. This will contribute to **20 marks**.

- Rubrics for each Experiment taken average for all Lab components – 15 Marks.
- Viva-Voce– 5 Marks (more emphasized on demonstration topics)

The sum of three tests, two assignments, and practical sessions will be out of 100 marks and will be **scaled down to 50 marks**
(to have a less stressed CIE, the portion of the syllabus should not be common /repeated for any of the methods of the CIE. Each method of CIE should have a different syllabus portion of the course).

CIE methods /question paper has to be designed to attain the different levels of Bloom's taxonomy as per the outcome defined for the course.

Semester End Examination:

Theory SEE will be conducted by University as per the scheduled timetable, with common question papers for the subject (**duration 03 hours)**

1. The question paper will have ten questions. Each question is set for 20 marks.
2. There will be 2 questions from each module. Each of the two questions under a module (with a maximum of 3 sub-questions), **should have a mix of topics** under that module.
3. The students have to answer 5 full questions, selecting one full question from each module
4. Marks scored shall be proportionally reduced to 50 marks

Suggested Learning Resources:

Text Books

1. Joel Grus, "Data Science from Scratch", 2nd Edition, O'Reilly Publications/Shroff Publishers and Distributors Pvt. Ltd., 2019. ISBN-13: 978-9352138326

Reference Books

1. Emily Robinson and Jacqueline Nolis, "Build a Career in Data Science", 1st Edition, Manning Publications, 2020. ISBN: 978-1617296246.
2. AurélienGéron, "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems", 2nd Edition, O'Reilly Publications/Shroff Publishers and Distributors Pvt. Ltd., 2019. ISBN-13: 978-1492032649.
3. François Chollet, "Deep Learning with Python", 1st Edition, Manning Publications, 2017. ISBN-13: 978-1617294433
4. Jeremy Howard and Sylvain Gugger, "Deep Learning for Coders with fastai and PyTorch", 1st Edition, O'Reilly Publications/Shroff Publishers and Distributors Pvt. Ltd., 2020. ISBN-13: 978-1492045526
5. Sebastian Raschka and Vahid Mirjalili, "Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2", 3rd Edition, Packt Publishing Limited, 2019.ISBN-13: 978-1789955750

Web links and Video Lectures (e-Resources):

1. Using Python : https://www.python.org
2. R Programming : https://www.r-project.org/
3. Python for Natural Language Processing : https://www.nltk.org/book/
4. Data set: https://bit.ly/2Lm75Ly
5. Data set: https://archive.ics.uci.edu/ml/datasets.html
6. Data set : www.kaggle.com/ruiromanini/mtcars
7. Pycharm : https://www.jetbrains.com/pycharm/

8. https://nptel.ac.in/courses/106/106/106106179/ 9. https://nptel.ac.in/courses/106/106/106106212/ 10. http://nlp-iiith.vlabs.ac.in/List%20of%20experiments.html
Activity Based Learning (Suggested Activities in Class)/ Practical Based learning 1. Real world problem solving - Applying the machine learning techniques and developing models